

The use of Diphone Variants in Optimal Text Selection for Finnish Unit Selection Speech Synthesis

Elina Helander, Hanna Silén, Moncef Gabbouj

Institute of Signal Processing, Tampere University of Technology, Finland

Abstract

The speech quality of a unit selection speech synthesizer depends highly on the database. This paper describes an approach for sentence selection for Finnish speech database recordings aiming at optimal coverage. The main idea is to define the diphone in a slightly different way: to distinguish diphones consisting of different allophones and also different linguistic positions, i.e. intra- and inter-syllabic diphones. We call these diphone variants. We evaluated if diphone variants become included in text selection for TTS prompt design without separate optimization and coarsely verified their acoustic dissimilarity. With the same number of sentences (292) that fulfill the traditionally determined diphone coverage completely, 66% more allophonic and inter/intra-syllabic contexts were missing with the conventional method compared to the proposed approach. We also describe how the approach inspired the synthesis process to reduce computational load.

1. Introduction

Unit selection [1] is a popular technique for implementing a text-to-speech (TTS) synthesizer. Unit selection based TTS systems utilize a large phonetically labeled speech database for choosing and concatenating segments in an optimal way. Optimal means that the synthesizer attempts to choose consecutive segments from appropriate contexts to avoid discontinuities and produce natural speech. The quality and naturalness that can be achieved surpasses the quality of traditional diphone-based techniques based on prosody modification.

A recent study on English TTS [2] showed that it is beneficial to separate pre- and post-vocalic consonants during synthesis. This separation could be implemented using more detailed target costs, which take contexts into account. However, if there are no good units available, the quality is degraded. Thus, the design of the inventory is important. Sentences for the inventory are usually selected automatically from a large collection of texts which saves time compared to manual design.

Covering all possible words or contexts is not possible for an open-domain TTS synthesizer and thus smaller units are optimized. A unit is usually a diphone or a triphone. In optimal text selection the aim is to cover the desired units with the smallest number of sentences. Greedy selection is a popular method applied to the optimal coverage problem and its advantage is significant if the size of the database is to be small [3]. The first sentence that becomes picked by the greedy algorithm has the largest number of different units. The sentence which maximizes the number of new units is the next one chosen. Here a new unit means a unit that is not yet present in the chosen sentences.

By optimizing only coverage the frequency of the units in a language is ignored. Some units appear much more often than others. The selection can also be carried out by taking into account the frequency of the units. Nevertheless, rare events are common in speech [4] and according to [5], using half-phones instead of natural rare diphones was not preferred. Thus it is important to include also rare units. Basically one should optimize all units in all phonetic and linguistic contexts which leads to a complicated sub-space problem with complex interactions [3]. Black and Lenzo [6] propose to search acoustically distinct units of a particular phoneme by building a classification and regression tree whose criterion is an acoustic distance measure between two units. This approach requires a speech database.

In this paper, we describe an approach for optimizing sentences with the greedy algorithm according to diphone and syllable coverage. The greedy selection is not frequency-weighted, since we are developing a rather small database that also contains rare units. Diphones are

defined in a slightly different way to account for allophones and syllable/word boundaries. No speech database is required but some linguistic knowledge about the allophones and syllabification of the language is required. Nevertheless, our approach avoids the complexity of the approach described in [3]. The purpose was to build a small unit selection speech database in Finnish concentrating on diphones, but the proposed idea can be extended to other languages with a high number of allophones and polysyllabic words, or to balance large databases, or to optimize triphone coverage. We built a speech database for unit selection synthesis from the variant aspect in diphones and in syllables. As mentioned, there was initially no speech database available to examine acoustically distinct units as in [6] but a database was recorded and a coarse evaluation was done afterwards.

The paper is organized as follows. Section 2 describes the motivation and idea of diphone variants. The process of building the database is described in Section 3. Analysis of the database with and without variants is provided in Section 4. In addition, acoustic evaluation using the proposed approach and how it motivated the synthesis are discussed. Section 5 concludes the paper.

2. Diphone variants as optimization units

For the prompt design for the speech database, allophonic and context-dependent variations of di-phones were explicitly included. This is particularly important for Finnish TTS systems due to the high number of allophones and polysyllabic words, and consequently inter-syllabic diphones. Some details of the Finnish language are provided in 2.1. The idea of diphone variants is described in 2.2.

2.1. Finnish language structure

Finnish orthography is phonemic: each phoneme corresponds to a certain grapheme with one exception (graphemes *ng* in *kangas* correspond to phoneme /ŋ/). A relatively high number of allophones exist due to the low number of consonants in Finnish. Most of the allophones are not pointed out in grapheme-to-phoneme conversion. Many consonants are articulated at a different place depending on the context, especially with front or back vowels. For example the phoneme /n/ has 5 allophones. Most of the consonants can also form geminates that are common. Contrary to the low number of consonants, there are rather many vowels. The vowels can appear as short or long and the quantity is distinctive. The differences between orthography and pronunciation mainly originate from boundary gemination [7]. In boundary gemination, a consonant at the beginning of word becomes geminated due to previous word ending with a vowel. The majority of Finnish words are polysyllabic and the syllable structure is simple with no complex consonant clusters.

2.2. Diphone variants

The starting point for the text database design was that no speech database was available. Thus our diphone variant based method (referred as to the *DV method*) does not have a way of acoustically determining distinct types. Since surrounding phonemes are relevant for the realization of phonemes and they are assumed to cause acoustical differences, the proposed approach takes into account how phonemes form different diphones in two cases:

- • The allophonic variants of a phoneme: e.g. the diphone *a_n* in the word *vanki* (prisoner) is considered different from the diphone *a_n* in the word *vanha* (old), due to allophonic variants of the phoneme /n/.
- • The linguistic position of a diphone: e.g. the diphone *a_n* in the word *vana* (*va-na*, trail in English) is considered different from *vanha* (*van-ha*), where - denotes the syllable boundary. Note that here the phonemes /n/ are not allophones.

If diphone variants are ignored, there is no guarantee that the database ends up containing all allophonic contexts and both inter-syllabic and intra-syllabic contexts if they exist. When the size of the database increases, it is more likely to contain the contexts not separately optimized. An example of the proposed transcription which separates the diphone variants is shown in Table 1. A number after a phoneme means an allophone of that phoneme and consonant geminates and long vowels are denoted by ":". The notation separates intra-syllabic (), inter-syllabic (-) and inter-word (- -) diphones. In syllable transcription, (*) denotes the primary stress and [*] denotes no stress, other syllables are not marked. The realization of allophones and syllabification in Finnish is obtained easily using hand-crafted rules.

Table 1. Transcription of a sentence with the conventional and the proposed method.

| |
|--|
| <p>Vanhemman veljen ansiosta nuorempi veli sai pilan anteeksi.</p> <p><i>Thanks to the older brother, the younger brother was forgiven the joke.</i></p> <p>Conventional</p> <p><i>Diphones:</i></p> <p>#_v v_a a-n n-h h-e e-m: m:_a a_n n_v v_e e_l l_j j_e e_n n_a a_n n_s s_i i_o o_s s_t t_a a_n n_u u_o</p> <p>o_r r_e e_m m_p p_i i_v v_e e_l l_i i_s s_a a_i i_p p_i i_l l_a a_n n_a a_n n_t t_e: e:_k k_s s_i i_#</p> <p><i>Syllables:</i></p> <p>van hem man vel jen an si os ta nuo rem pi ve li sai pi lan an te:k si</p> <p>Proposed</p> <p><i>Diphones:</i></p> <p>#_v v_a a_n1 n1-h h_e e_m: m:_a a_n1 n1--v v_e e_l3 l3-j j_e e_n1 n1--a a_n1 n1-s s_i i_o o-s s_t</p> <p>t_a a--n1 n1_u u_o o-r r_e e_m1 m1-p p_i i--v v_e e-l3 l3-i i--s s_a a_i i--p p_i i-l2 l2_a a_n n--a</p> <p>a_n2 n2-t t_e: e:_k1 k1-s s_i i_ #</p> <p><i>Syllables:</i></p> <p>(van) hem [man] (vel) [jen1] (an1) si os [ta] (n1uo) rem [pi] (ve) [li] sai (pi) [lan1] (an2) te:k [si]</p> |
|--|

3. Database construction and statistics

Before text optimization, phonetization and spelling rules for a language must be defined. In case of Finnish they are rather simple excluding foreign and some compound words. Simple punctuation rules were used for marking pauses and a pause was considered as a part of a diphone as well. A geminate consonant was modeled as a phoneme separately from single consonants. A diphone in a word boundary prone to boundary gemination was ignored since its realization in the read speech is not consistent. A diphone combining two words where the last word starts with a vowel is used in optimization.

Following the idea of CMU Arctic database [8], texts were derived from out-of-copyright books. In total 33 Finnish books with 203 339 sentences from Project Gutenberg [9] were extracted. The sentences containing 6-15 words were selected and the resulting set of 46 067 sentences was used in the optimization process referred to as the source data. Less than 17 % of the words in the source set were monosyllabic leading to a relatively high amount of inter-syllabic diphones. This supports the idea of separate optimization of inter- and intra-syllabic

diphones. For comparison, about 72 % of words in the 1032 utterances of the English CMU Arctic data [8] are monosyllabic.

The text selection process was done in two phases: first a set with full diphone variant coverage was built (referred to as Set A) resulting in 424 sentences. Then a second set was built to optimize syllable variants (Set B). Since the aim was to build a rather small database, 600 sentences were chosen for Set B. After manual pruning, the database contained 1003 sentences.

The purpose of Set A was to cover all diphone variants in Finnish. Table 2 summarizes the number of different diphones encountered in the sentence set with and without considering diphone variants. The percentage of diphones occurring once or twice is slightly less without diphone variants.

Table 2. The number of diphones/diphone variants and rare diphones/diphone variants in 46 067 sentences.

| | No variants | With variants |
|-------------------------------|-------------|---------------|
| Number of units | 1112 | 1585 |
| Units occurring once or twice | 12.1 | 12.4 |

Set B was designed to be rich in different syllables. Since the main stress in Finnish is always on the first syllable and the last syllable is always unstressed, both of these contexts were separately included. For example in Table 1, syllable *pi* as stressed in *pila* is now optimized separately from unstressed ones (i.e. in *nuorempi*). In addition, syllables were determined with allophones, i.e. for example in Table 1 the syllable *an1* is considered different from syllable *an2*. However, the effect of most of the allophones (e.g. allophones of /k/ and //) remains inside a syllable and do not need to be marked.

Syllable variants already included in Sentence A were taken into account. The obtained syllable variant coverage is shown in Table 3. Since the use of creaky voice at the end of a sentence is a frequent phenomenon in Finnish [10], the last word was not used in the optimization. The first word was also ignored as well as monosyllabic words whose stress pattern differs from polysyllabic words.

Table 3. The number of syllable variants in Set A and Set B versus the source data.

| | Stressed | Unstressed |
|---------------|----------|------------|
| Source data | 2345 | 1981 |
| Set A + Set B | 1695 | 1466 |

4. Evaluation

A speech database of 1003 sentences resulting from the optimization process described in section 3 was recorded. The sentences were recorded by a female voice at a sampling frequency of 32 kHz. For the alignment, HMM-based phoneme models were trained and sentences were forced-aligned with the phoneme transcription.

The evaluation of the proposed method is not straightforward. Evaluation through recording two different databases is not practical. We carried out experiments on textual coverage, acoustic similarity between diphone variants in the speech database, and inter- and intra-syllabic diphone pre-selection in synthesis.

4.1. Diphone variant coverage in Set A

We analyzed how traditional selection (referred as to the NV method) which does not consider diphone variants succeeds in including them without separate optimization. Both the NV and the DV methods utilized the greedy algorithm to select the sentences until no diphones/diphone variants were missing. The NV method selected 292 sentences to cover all the required diphones. For the DV method it took more, 424 sentences, since there were more units to be covered. After 144 sentences the NV method added only one new diphone while the respective number for the DV method was 191. Further, since the number of sentences required for the total coverage is naturally smaller with the NV method, only the first 292 sentences of the DV method were used for evaluation.

We examined how many diphone variants were missing within those 292 sentences chosen by the both methods. The NV method had 219 diphone variants missing (13.8%), although it had all the conventional diphones covered. The DV method had 132 diphones missing (8.3%) with 292 sentences. Furthermore, we examined the missing diphone variants for both methods as a function of the number of sentences by calculating the coverage after each added sentence. The results are shown in Figure 1. Naturally the DV method performs better since variants are its optimization criteria but the figure rather illustrates that diphone variants do not become randomly picked along with the NV method.

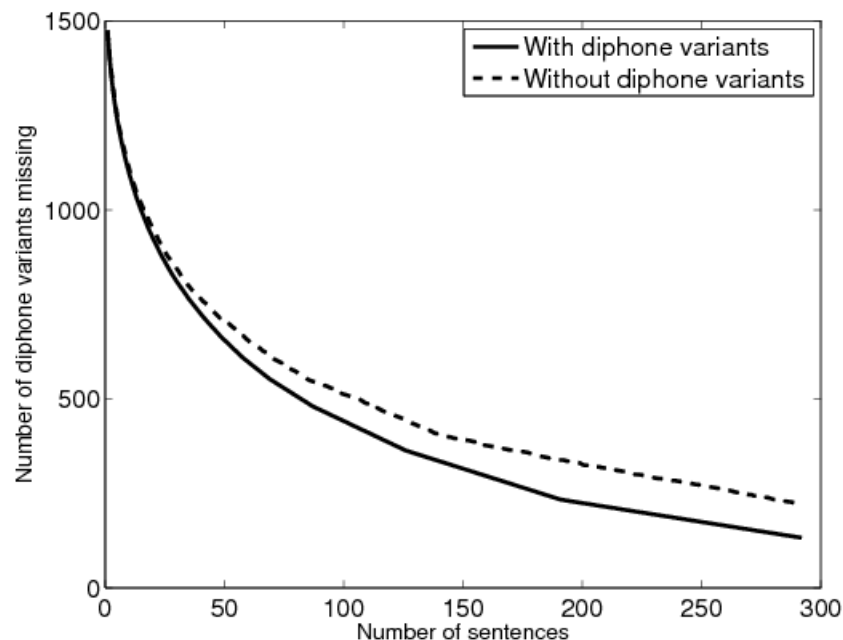


Figure 1. Number of diphone variants missing with (the proposed DV method, solid line) and without (the NV method, dashed line) separate optimization.

4.2. Acoustic evaluation

We determined acoustic distances between diphone variants that are traditionally considered the same. Although acoustic distances are database- and speaker-specific [6], some coarse guidelines can be obtained. Acoustic distance based on 13 normalized mel-frequency cepstral coefficients (MFCC) was calculated at diphone level. The idea is adopted from [6] with slight modification for taking into account the diphone boundary and already normalized values. The acoustic distance between unit U and unit V is defined in two parts where U_1 and V_1 are the first parts of diphones U and V consisting of N_1 and M_1 MFCC frames, respectively. The last part of the diphones U and V are denoted by U_2 and V_2 , respectively, and the lengths by N_2 and M_2 . The total acoustic distance is a sum of the distances between both pairs:

$$d(U, V) = d(U_1) + d(U_2) \quad (1)$$

where

$$d(U_k, V_k) = \beta \cdot \frac{1}{L_k} \sum_{i=1}^{L_k} \sum_{j=0}^{12} |c_j(i) - c_j(y(i))| \quad (2)$$

where $k=1,2$; $L_k=\max\{M_k, N_k\}$; $c_j(i)$ denotes the j^{th} normalized MFCC coefficient of frame i of the longer unit and $y(i)$ is the corresponding frame in the shorter unit:

$$y(i) = \left[i \cdot \frac{\max\{M_k, N_k\}}{\min\{M_k, N_k\}} \right] \quad (3)$$

where $[*]$ denotes nearest integer rounding. The factor β in (2) denotes the duration penalty for the acoustic distance and is defined as:

$$\beta = \alpha \cdot \frac{\max\{M_k, N_k\}}{\min\{M_k, N_k\}} \quad (4)$$

where α is a weighting factor for the duration ratio difference.

Consider diphone variants d_1 and d_2 that are traditionally defined as the same diphone. The two cases of a diphone variant are defined in 2.2. Instances of d_1 and d_2 form classes c_1 and c_2 , respectively. Now we calculate intra-class acoustic distances between all class members in class c_1 (or c_2) and compare them to inter-class distances. If there are m instances of d_1 and n instances of d_2 , there are m_2-n diphone variant distances between the members of c_1 , m_2-m of c_2 and $n \cdot m$ inter-class distances between the members of c_1 and c_2 . For example for diphone $e_/_$ differences between all intra-syllabic instances (e.g. *veljen*, Table 1) are calculated. The same procedure is repeated for each inter-syllabic instance $e_/_$ (e.g. *velan*, Table 1). Finally, the distance between every intra- and inter-syllabic instance of $e_/_$ is calculated. For each diphone variant pair, intra- and inter-class distances were compared using the two-tailed t-test with hypothesis of equal means at 5% significance level.

Since there can be only a few of some diphone variants and statistical reliability would be rather low, we only consider diphone variant pairs that have at least 20 instances per each. In total 54 pairs were used for evaluation. Every instance was checked manually and erroneous instances were discarded. The summary of t-test results is shown in Table 4. In 54% of the cases both intra-class distance means were significantly lower than inter-class distance mean and in 85% one or both intra-class distance means were significantly lower. Note that only significant mean differences are considered, in many cases the intra-class distance mean was lower than inter-class distance mean, but not significantly. Duration penalty factor in (4) was set to 1, since the value did not substantially affect the results.

Table 4. The comparison of intra- and inter-class distances

| | |
|---|----|
| Both intra-class means significantly higher | 29 |
| One intra-class mean significantly lower, the other equal | 17 |
| Both means equal to inter-class means | 3 |
| One intra-class mean significantly lower | 5 |
| Both intra-class means significantly higher | 0 |
| Number of pairs in total | 54 |